DOCUMENT RESUME

ED 442 857 TM 031 277

AUTHOR Chiu, Christopher W. T.

TITLE A Subdividing Method for Generalizability Theory: Precision of Measurement Errors and Patterns of Missing Data.

PUB DATE 2000-04-00

NOTE 43p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000). This study was based on a dissertation (Chiu, 1999) funded by the Dissertation Awards from the Graduate School at Michigan State University (MSU) and the Society of Multivariate Experimental Psychology (SMEP). The study equipment was sponsored by the Educational Testing Service (contract number: Ref. No. 5530) and by the computer center at MSU.

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.

DESCRIPTORS Difficulty Level; *Error of Measurement; *Generalizability Theory; Interrater Reliability; Monte Carlo Methods; Simulation

IDENTIFIERS *Missing Data

ABSTRACT

A procedure was developed to analyze data with missing observations by extracting data from a sparsely filled data matrix into analyzable smaller subsets of data. This subdividing method, based on the conceptual framework of meta-analysis, was accomplished by creating data sets that exhibit structural designs and then pooling variance components from these designs. A Monte Carlo simulation was used to examine the statistical properties of the variance-component estimates and some commonly used composite indices, the generalizability coefficient and the dependability coefficient. Data sets used to evaluate the method ranged from 750 examinees and 4 raters to 6,000 examinees and 28 rates with 2 tasks. Experimental conditions, such as item difficulty and rater inconsistency, were varied to model operational performance assessments. The subdividing method recovered variance component estimates with high accuracy and precision in a variety of conditions. Increasing the number of examinees scored by the same raters from 12 to 24 had virtually no effect on the accuracy and precision of the estimates. Results indicate that the subdividing method produces outcomes having properties that are similar to those of complete data methods. Different rules for forming groups of rates changed the structural design of scores and thus influenced the precision of measurement error estimations. Results from this study indicate that scoring centers can determine and forecast how confidently they can interpret generalizability analyses by controlling the rules used to assign tasks to raters and the universe of generalization for which normative and criterion-referenced decisions are made. Appendixes contain discussions of the structure of a modified balanced incomplete block design, a model to determine the size of a rater pool, and equations for the standard error of variance components in a two-facet crossed design. (Contains 13 figures and 31 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running Head: A subdividing method and missing data

ED 442 857

A Subdividing Method for Generalizability Theory: Precision of Measurement Errors and Patterns of Missing Data

Christopher W.T. Chiu

Program in Research Methodology

Department of Psychology in Education
School of Education

University of Pittsburgh

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Cchiu

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Note

Paper to be presented at the 2000 NCME Annual Conference (April 24-28, 2000, New Orleans, LA) sponsored by the National Council on Measurement in Education. Correspondence concerning this article should be addressed to Christopher W.T. Chiu, 5C24 Forbes Quad, University of Pittsburgh, Pittsburgh, PA 15260. Electronic mail may be sent via Internet to cchiu+@pitt.edu. Telephone: (412) 605-0367. Christopher W.T. Chiu specializes in Measurement and Psychometric Theory. This study was based on a dissertation (Chiu, 1999) funded by the Dissertation Awards from the Graduate School at Michigan State University and the Society of Multivariate Experimental Psychology (SMEP). The simulation in this study would not have been possible without the equipment sponsored by the Educational Testing Service (contract number: Ref. No, 5530) and by the computer center at Michigan State University. The author thanks Betsy Becker, Robert Floden, Michael Harwell, Clement Stone, Niel Timm, Edward Wolfe for their comments and suggestions.

TM031277

BEST COPY AVAILABLE

Running Head: A subdividing method and missing data

A Subdividing Method for Generalizability Theory: Precision of Measurement Errors and Patterns of Missing Data

Christopher W.T. Chiu

Program in Research Methodology

Department of Psychology in Education
School of Education

University of Pittsburgh

Abstract

In generalizability analyses, unstable and potentially invalid variance component estimates may result from using only a limited portion of available data. However, missing observations are common in operational performance assessment settings (e.g., Brennan, 1992 and 1997; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) because of the nature of the assessment design. In this study, I described a procedure to analyze data with missing observations by extracting data from a sparsely-filled data matrix into analyzable smaller subsets of data. This subdividing method, drawing on the conceptual framework in meta-analysis (e.g., Hedges & Olkin, 1985), was accomplished by creating data sets that exhibit structural designs (i.e., *crossed*, *nested*, and *modified balanced incomplete block* designs) then pooling variance components obtained from these designs. This method is always more computationally effective than other methods that require sparsely collected scores to be analyzed all at once. A Monte Carlo simulation was used to examine the statistical properties of the variance-component estimates and some commonly used composite indices, namely the generalizability coefficient (for norm-referenced decisions) and the dependability coefficient (for criterion-referenced decisions). The smallest unbalanced data set used to evaluate the subdividing method was composed of 750 examinees, four raters, and two tasks while the largest unbalanced data subset was composed of 6000 examinees, 28 raters, and two tasks. Graphic displays were used to evaluate the accuracy, stability, and consistency of the variance component estimates and the composite indices. Experimental conditions, modeling operational performance assessments, were manipulated to examine how well the subdividing method would perform in practice. These conditions included: (1) volume of examinees, (2) size of rater pool, (3) variation in item difficulty, (4) levels of rater inconsistency, (5) rules used to decide how to group raters and assign tasks to raters, and (6) the minimum number of examinees scored by a group of raters.

The subdividing method recovered variance component estimates with high accuracy and precision in a variety of conditions (i.e., low and high variations in item difficulty and rater inconsistency). Increasing the number of examinees scored by the same group of raters from 12 to 24 had virtually no effect on the accuracy and precision of the variance component estimates. Results indicated that the subdividing method produced outcomes having properties (unbaisedness and consistency) that are similar to those of complete data methods. For instance, the larger the volume of the examinees, the more precisely the following effects were estimated: *person*, *person-by-item*, *person-by-rater*, and *person-by-item-by-rater* effects. In addition, even though every essay was scored by only two randomly selected raters from a large pool of raters, increasing the size of the pool of raters improved the precision in determining rater-related effects (i.e., *rater*, *item-by-rater*, *rater-nested-in-item* effects). Different rules used for forming groups of raters changed the structural design of scores and thus influenced the precision of measurement error estimation. The results from a Monte Carlo study concluded that scoring centers can determine and forecast how confidently they can interpret generalizability analyses by controlling (1) the rules used to assign tasks to raters and (2) the universe of generalization for which normative and criterion-referenced decisions are made.

Introduction and Conceptual Framework

In generalizability analyses, unstable and potentially invalid variance component estimates may result from using only a limited portion of available data. However, missing observations are common in operational assessment settings (e.g., Brennan, 1992 and 1997; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) because of the nature of the assessment design. For instance, missing observations can be a result of the use of different groups of raters to score different sets of examinees (e.g., Brennan, Gao, and Colton, 1995). In this study, I describe a procedure to analyze data with missing observations by extracting data from a sparsely-filled data matrix into analyzable smaller subsets of data. This subdividing method, drawing on the conceptual framework in meta-analysis (e.g., Hedges & Olkin, 1985), is accomplished by creating data sets that exhibit structural designs (i.e., *crossed*, *nested*, and *modified balanced incomplete block* designs) then pooling variance components obtained from these designs. Brennan (1992) illustrates the *crossed* and *nested* designs and Appendix A shows the *modified balanced incomplete block* design. The subdividing method is always more computationally effective than methods (i.e., Maximum Likelihood, Restricted Maximum Likelihood, Minimum Norm Quadratic Unbiased Estimate) that require sparsely collected scores to be analyzed all at once.

Problem Statement and Research Questions

Currently, researchers are frequently forced to discard data when some data are missing in large-scale assessments, leading to unstable variance components and reliability coefficient estimates used to evaluate measurement procedures. Interpreting those unstable estimates can lead to inconsistent decisions (Burdick 1992, pp.16-18). For example, if our goal is to develop a performance assessment procedure with a generalizability coefficient of 0.90, we may reach different conclusions using the following three confidence intervals [0.85, 0.89], [0.44, 0.89], and [0.44, 0.50]. In the first case, we may conclude that for practical purposes, the generalizability coefficient is close enough to 0.90 that it may not be efficient to increase the reliability. In the second case, we may decide that the confidence interval is too wide to make

conclusive decisions about the reliability of the assessment procedure. In the final scenario, the evidence suggests that the assessment procedure is not well developed. In the context of scoring performance tasks in which observations and judgements are involved, the final scenario can indicate that raters differ greatly in severity and need further training. Missing data influence the precision of the variance estimates and thus introduce noise in making high-stake decisions. The following research questions tap into the statistical properties of the subdividing method as well as practical issues that are of the concerns of testing companies when different groups of raters score different sets of examinees.

Q1. How did the amounts and patterns of missing data influence the accuracy and precision of the variance components?

1a) Amounts of missing data: Changing the size of a hypothetical rater pool while holding constant the number of times a task was rated changed the amount of missing data. If five examinees were crossed with four raters, the number of possible ratings per item is 20 (5 examinees x 4 raters). However, if one randomly chooses only a pair of raters out of the pool of four raters, the number of possible ratings reduces to 10 (5 x 2), which is 50% of the total available ratings if all four raters were used (20). The other half of the ratings would be missing because they were unobserved by design. The percent of unobserved data increases further as the rater pool size expands to six raters. Here using a pair of raters resulted in 66.6% of all possible data being unobserved (10 ratings of 30 possible are observed). The proportion of missing data increases as the size of the rater pool increases, holding constant the number of ratings each examinee received. The current study manipulated the rater pool size to investigate the effects of amounts of missing data on the accuracy and precision of the subdividing method.

1b) Patterns of missing data: Although data collection procedures are critical, rarely did research set out to investigate how these procedures influence the statistical properties of G theory estimates (Personal communication from Gordon, 1998; Vickers, 1998; & Welch, 1996). The current study summarizes two data collection procedures frequently appeared in the literature for scoring open-ended questions based on human judgements (Brennan, Gao, & Colton, 1995; Chiu & Wolfe, 1997; Lane, Liu,

Ankenmann, and Stone, 1996; and Linn et al., 1996). Data collection procedures determine the pattern of missing data (Engelhard, 1997) and thus influence the precision of parameter estimates (Little & Rubin, 1987). In multiple facet generalizability studies employed for large-scale assessments, the ways to assign tasks to raters and the mechanisms to distribute examinees' work to raters constitute the data collection procedures, also defined as "*rating plans*" throughout the current study. A collection of rules, also defined as a *rating plan*, was examined. Specifically, this study compared two prevalent rating plans (i.e., the *disconnected crossed* and *connected mixture* plans). The robustness and performance of the proposed subdividing method is evaluated in operational settings parallel to the two data collection procedures. The data collection procedures differ in the *grouping of raters* and the *assignments of tasks to raters*. These two factors are discussed in the following.

i. Grouping of raters: some scoring centers divide raters into distinct pairs and assign different sets of examinees to the pairs of raters and thus rater groups do not share raters (i.e., each rater belongs to one and only one group). However, other scoring centers do not have this restriction in grouping raters and consequently rater groups share raters (e.g., each rater can belong to more than one group).

ii. Task assignments: scoring centers can either assign individual essays or intact portfolios to raters. Patz, Wilson, and Hoskens (1997) define an intact portfolios or booklet as a set of essays submitted by an examinee.

A rating plan refers to the ways to assign tasks to raters and the mechanisms to distribute examinees' work to raters. Two aforementioned rating plans were manipulated (*disconnected crossed* and *connected mixture* rating plans). Chiu and Wolfe (1997) called the former and latter plans rating plans with high and low standardization, respectively. The *disconnected crossed* rating plan was one in which specific rules were used to guide the assignment of tasks whereas in the *connected mixture* plan tasks were routed to raters completely at random.

In practice, different versions of the two rating plans introduced in the above section were adopted in various scoring centers (Gordon, 1998; Vickers, 1998) and the principles shown in Table 1 (p. 6) capture the essentials.

Table 1: Principles of rating plans

Principles	Disconnected crossed plan	Connected mixture plan
1. Number of ratings on each task for a given examinee	2	2
2. Linking between groups	raters belong to only one group	no restriction
3. Guiding structural designs within a given group	limited to only the crossed design	no limitation; structural design could be crossed, nested, and other designs
4. Number of examinees scored by groups of raters.	planned; every group of raters scores the same number of examinees.	unplanned; groups may or may not score the same number of examinees depending on the size of rater pool and sample size.

Q2. Can the subdividing method handle a large volume of examinees? How well does it perform?

Large-scale assessments can have numbers of examinees ranging from a few hundred to several thousand, or even tens of thousands for state and national tests. Longford (1995) reported that 3,756 examinees responded to the Studio Art Portfolio Assessment and Myford, Marr, and Linacre (1995) reported 5,400 examinees took the Test of Written English (TWE) in one administration. Chiu and Wolfe (1997) stated that 5,905 examinees participated in an administration of the Collegiate Assessment of Academic Proficiency (CAAP). Lane et al. (1996) conducted generalizability analyses on 2,514 examinees who had responded to all the tasks in the QUASAR Cognitive Assessment Instrument (QCAI). How well can the subdividing method perform for large-scale assessments?

Q3. To what degree does increasing the number of essays scored by a common group of raters increase the precision of the variance component estimates and the reliability coefficient estimates?

Other than randomly assigning essays to raters, scoring centers (e.g., Georgia State Department of Education and ACT, Inc.) often impose rules for the scoring procedures to accommodate operational needs. An important dimension of setting up a scoring procedure is to arrange essays so that they can be efficiently graded. For example, rather than randomly assigning every essay to each rater, scoring centers

often organize essays in batches (Gordon, 1998; Schafer, 1998; Vickers, 1998; Welch, 1996; Wolfe, 1998). Those batches can be randomly assigned to groups of raters. Packing essays in batches saves operational time because it takes more time for raters to exchange single essays than a batch of essays. Packing essays in bundles also controls the number of essays to be scored by a common group of raters. Bundling also structures ratings for reliability analysis — without the bundling of essays, the data set may be too sparsely-filled to conduct an analysis.

Q4. When should rating plans with high standardization be used in placed of rating plans with low standardization?

Chiu (1999) found that the two factors, *Grouping of Raters* and *Task Assignment* (discussed on page 5) together influenced the amount of data available to estimate the rater-related effects. Specifically, a rating plan with low standardization will have too few data points to estimate rater-related effects separately. As a result, one cannot to disentangle $\hat{\sigma}^2_r$ from $\hat{\sigma}^2_{re}$ and $\hat{\sigma}^2_p$ from $\hat{\sigma}^2_{pr,e}$ (Brennan, 1992). This finding suggested that the degree of standardization in assigning tasks to raters was related to the amounts of data used for estimation and thus should affect the precision of the estimates of variance components. If a low degree of standardization is employed to collect data, how much precision will be compromised? In addition, when should a highly standardized rating plan be used? (i.e., in what situation should raters be designated to work as independent groups which score intact portfolios?)

Methodology

This section summarizes the procedures of the subdividing method and describe a Monte Carlo simulation to evaluate the subdividing method. Monte Carlo studies (Cronbach et al., 1972; Mooney, 1997) are especially suitable for the current study because one can evaluate the practicality and statistical properties (e.g., bias and consistency) of an estimation method by comparing the estimated parameters to the known population parameters. In the measurement context, it is infeasible to conduct many reliability experiments in which actual raters and examinees are crossed. Also, it is difficult to imagine how one

could control precisely raters' severity and inconsistency for such experiments. In addition, the real problem is that one cannot evaluate estimation methods based on data for which one does not know the population parameters. These restrictions make Monte Carlo studies especially appropriate for examining this subdividing method. Furthermore, as has been pointed out by many researchers (e.g., Harwell, Stone, Hsu, & Kirisci, 1996; Longford, 1995; Psychometrika Editorial Board, 1979), analytical methods (e.g., deriving expected mean squares equations) can be inadequate for the examination of statistical properties in generalizability analysis. Unlike analytical methods, simulation studies can be used even if probabilities of selection (determined by the amount of missing data), sample sizes, and the magnitude of variance components are treated as independent variables. Also, Monte Carlo methods are especially suitable when it is difficult to satisfy asymptotic assumptions because only a small number of levels are sampled in each factors (e.g., only two levels in the item facets, $n_i = 2$).

Procedures of the subdividing method

The subdividing method had three stages. They were the 1) Modeling, 2) Estimating, and 3) Synthesizing stages. The sections that follow summarize each of these stages, which were discussed in detail in Chiu (1999).

Modeling Stage — In the first step, the sparsely filled data set was divided into smaller subsets of balanced data that exhibited structural designs common in Analysis of Variance, namely the *crossed* design, the *nested* design, and the *modified balanced incomplete block design* (*MBIB*). The sparsely-filled data set was divided into S_t data subsets, with $t = 1, 2,$ or 3, and with S_t indicating the number of subsets in each of the *crossed* (S_1), *nested* (S_2), and *MBIB* (S_3) designs, respectively. Note that the *crossed* and *nested* designs were structural designs that are common in generalizability analysis. A *MBIB* design was formed every time one rater scored both items and was paired with a different second rater on each item.

Estimating Stage — In this stage, variance components were estimated for each subset of data. The ANOVA method (e.g., Brennan, 1992, Searle et al., 1992, p. 173) was used to estimate variance components for the *crossed* and *nested* designs. For the *MBIB* design, the Minimum Norm Quadratic

Unbiased Estimate (MINQUE) method was used to obtain the variance component estimators (Bell, 1985; Giesbrecht, 1983; Goodnight, 1978; Rao, 1997).

Synthesizing Stage — Meta-analysis (Hedges and Olkin, 1985), a quantitative method to summarize research results, is especially suitable for the subdividing method because it is capable of estimating an overall outcome based on many outcomes obtained from individual empirical studies, or, here, data subsets. Thus, meta-analytic methods were used to aggregate variance components from each subset of data. In the Synthesis Stage, data subsets were weighted by subset sample size and then variance components were pooled. Composite indices were computed on the basis of those pooled variance component estimates.

Weighted estimates of the variance components were obtained by weighting the data subsets by their sample sizes across all subsets from both the *disconnected crossed* and *connected mixture* rating plans. For factor *f* we obtained

$$\hat{\sigma}^2_f = \bar{\sigma}^2_f = \frac{\sum_{t=1}^{T}\sum_{s=1}^{S_t} n_{p,i,s}\,\hat{\sigma}^2_{f,s}}{\sum_{t=1}^{T}\sum_{s=1}^{S_t} n_{p,i,s}}, \qquad (1)$$

where $f = \begin{cases} p, i, r, pi, pr, ir, \text{ and } pir & \text{for a } \textit{crossed} \text{ or } \textit{MBIB} \text{ design} \\ p, i, r{:}i, pi, \text{ and } p{:}ri & \text{for a } \textit{nested} \text{ design} \end{cases}$

s = the s^{th} data subset

t = the t^{th} structural design, and

$n_{p,i,s}$ = number of examinees in the s^{th} data subset of the t^{th} structural design.

When the data subsets were equal in sample size (i.e., had the same number of examinees), the weighted average variance component became:

$$\hat{\sigma}_f^2 = \overline{\sigma}_f^2 = \frac{\sum_{s=1}^{S} \hat{\sigma}_{f,s}^2}{S.}, \qquad (2)$$

where $S.$ is the number of data subsets across all the structual designs.

Conditions to vary

Rating plans (2 levels): The aforementioned *disconnected crossed* and the *connected mixture* rating plans were examined in the simulation study.

Number of examinees (4 levels): Four levels were used in the current study: 750, 1500, 3000, and 6000 to reflect the realistic size of examinees tested in large-scale setting. These four levels were chosen by modeling the sample size in large-scale testing programs (Chiu and Wolfe, 1997; Lane et al., 1996; Longford, 1995; and Myford, Marr, and Linacre, 1995).

Number of raters (3 levels): The size of the rater pool was varied to reflect practical situations. In the study by Lane et al. (1996), 34 raters were hired to score the QCAI examinations of 2,514 examinees. In another study by Chiu and Wolfe (1997), nine raters were used to score 5,905 examinees. The two examples (Lane et al., 1996 and Chiu & Wolfe, 1997) showed that the number of examinees and the size of the pool of raters needed not be in direct proportion. Many other intervening operational factors influenced this functional relationship. Such factors included the *number of tasks answered by an examinee*, *number of ratings on each task*, *total number of days available for scoring*, *time (in minutes) it takes to score a task*, and *average work hours per rater per day*. Appendix B in the study showed an equation to determine the number of raters needed to complete the scoring of an examination for varying sample sizes, while holding constant the other operational factors. Three levels were used for the 750-, 1500-, and 3000- examinee conditions and 2 levels for the 6000 examinee condition.

Number of items and variation in item difficulty (2 levels): Data were simulated to model a case with two items and two ratings per item. This choice reflected the common practice in examinations where essay writing was involved (e.g., Collegiate Assessment of Academic Proficiency or CAAP,

Graduate Management Admission Test or GMAT, and Medical College Admission Test or MCAT). Each item was scored two times (occasions) by completely different raters. Variations in item difficulty were chosen to reflect the variations reported in the literature. Table 2 showed a summary of the variance components reported in four published studies involving human judgments. The third column indicates the scoring scale employed in the studies. As the studies employed a different scoring scale (a 6-point scale on the first two studies and 5-point scale on the last two), it was necessary to use a common metric (relative percent of variation based on the total variation) to compare the variance components. The average percentage were used as the population parameters in the simulation study.

Table 2: Summary of variance component magnitudes in the literature

Authors	Year	Subjects	Scoring Scale	Total Variation	Person P	Item I	Rater R	PI	PR	IR	PIR, E
Brennan et al	1995	Writing	0 to 5	1.15967	59%	2%	1%	14%	4%	1%	19%
Chiu et al	1997	Writing	1 to 6	0.47281	41%	5%	0%	25%	4%	1%	23%
Lane et al	1996	Math	0 to 4	1.84215	25%	11%	0%	53%	0%	0%	10%
Linn et al	1996	Math	0 to 4	1.01500	20%	22%	0%	33%	1%	2%	21%
			Average	0.96331	35%	11%	0%	31%	3%	1%	20%

Rater inconsistency (2 levels): Longford (1995, pages 21-2) defined σ^2_{pr} as the *rater inconsistency*. *Low rater inconsistency* was defined as $\sigma^2_{pr} = 0.01$ and *high rater consistency* was defined as $\sigma^2_{pr} = 0.1$.

Number of essays in a batch (2 levels): The two levels of batch size were 12 and 24, respectively. These levels reflected batch sizes being used in large-scale testing programs (Gordon, 1998; Schafer, 1998; Vickers, 1998; Welch, 1996; Wolfe, 1998).

Evaluation of the subdividing method (outcomes and data analysis)

Accuracy and Precision — The *Accuracy* and *Standard Error* measures (Chiu, 1999; Othman, 1995), derived from Mean Square Error (MSE) (Harwell et al., 1996 & Othman, 1995) were used to evaluate the subdividing method. The *Accuracy* index reflected the degree to which the subdividing method can recover the population parameters. *Accuracy* equals 1.0 shows that the subdividing method

BEST COPY AVAILABLE

recovers the parameters with 100% accuracy. The *Standard Error* measure is the standard deviation of the recovered parameter values and it has a range from 0 to infinity. The lower this index, the higher the precision.

Empirical versus Theoretical Standard Errors (SEs) — Standard errors based on asymptotic assumptions (Brennan, 1992, pp. 133-135; Burdick & Graybill, 1992) can be inaccurate as the degree to which the SE reflects the sampling distribution of a variance component depends upon factors like sample size, normality, and the amounts and patterns of missing data involved. According to Brennan (1992), given that the asymptotic assumptions are satisfied, the theoretical SEs for the variance components are functions of the mean squares of the facets and their degrees of freedom. The theoretical standard errors reported in the Results section were based on the formulas in Appendix C of the study, computed using two raters. Two raters were chosen because only a random pair of raters was selected to score each examinee, even though more than two raters were available in the pool.

Ratio of Standard Errors — The ratio between the standard errors of the variance components were used to evaluate the effect of packing essays into batches of different sizes. Ratio of standard errors equals $SE(\hat{\sigma}'^2_a)/SE(\hat{\sigma}^2_a)$, where $\hat{\sigma}'^2_a$ and $\hat{\sigma}^2_a$ represent the variance components or composites obtained by using a batch size of 24 and 12, respectively. A ratio greater than one indicates that the standard errors of the variance component obtained using a batch size of 24 was greater than that of using a batch size of 12.

Table 3 (p. 13) summarizes the conditions used to evaluate performance of the subdividing method. Each of the 176 conditions was replicated with 100 trials. Data were generated using a two-faceted linear model exhibiting a crossed structure (Schroeder, 1986, p.36):

$$X_{pir} = x_p + x_i + x_r + x_{pi} + x_{pr} + x_{ir} + x_{pir,e} \qquad (3)$$

Each of the seven components were generated under a normal distribution $x_a \sim (0, \sigma^2_a)$, where $\sigma^2_a = \{\sigma^2_p = 0.3;\ \sigma^2_i = 0.02$ (low), 0.11 (high); $\sigma^2_r = 0.01;\ \sigma^2_{pi} = 0.30;\ \sigma^2_{pr} = 0.01$ (low), 0.1 (high); $\sigma^2_{ir} =$

0.01; and $\sigma^2_{pir,e}$=0.20). For example, the score for person *p*, responding to item *i*, judged by rater *r*, was the sum of seven random numbers each generated independently from the above seven normal distributions. These population values of the variance components were chosen in coherence with the values reported in the literature (i.e., Brennan, Gao, & Colton, 1995; Chiu & Wolfe, 1997; Lane et al., 1996; Linn et al., 1996).

Table 3: Experimental conditions to evaluate the subdividing method

Experimental Conditions			High Item Effect (σ^2_i)				Low Item Effect (σ^2_i)			
			High rater inconsistency (σ^2_{pir})		Low rater inconsistency (σ^2_{pir})		High rater inconsistency (σ^2_{pir})		Low rater inconsistency (σ^2_{pir})	
	N examinees	N raters	12 essays per batch	24 essays per batch	12 essays per batch	24 essays per batch	12 essays per batch	24 essays per batch	12 essays per batch	24 essays per batch
Disconnected Crossed Rating Plans	750	2								
		4								
		8								
	1500	4								
		8								
		14								
	3000	8								
		14								
		28								
	6000	14								
		28								
Mixture of Connected Rating Plans	750	2								
		4								
		8								
	1500	4								
		8								
		14								
	3000	8								
		14								
		28								
	6000	14								
		28								

Note: The numbers of raters reported indicate the number of raters needed to complete the scoring session in 40, 20, and 10 days, respectively. It is assumed that examinees respond to two items and each item is scored twice by two different raters and it takes 10 minutes to read an essay and all raters work 7.5 hours a day. Consequently, it will take 40 days to score when a low volume of raters (i.e., 2, 4, 8,& 14) is recruited for each N examinees value. The scoring time will decrease by 50% when a medium rater pool is employed (i.e., 4, 8, 14, & 28). It takes only 10 days when a large pool of raters is recruited (i.e., 8, 14, & 28). A rater pool size of 4 is considered medium when it is used to score 750 examinees but it is considered low when is used to score 1500 examinees. The same rule applies to other size of rater pools.

BEST COPY AVAILABLE

Results

Two major findingsconcerning the accuracy and precision of the variance components estimated by the subdividing method were discussed in subsequent sections and they were summarized in below. First, the variance components for both the *disconnected crossed* and *connected mixture* rating plans (defined on page 5) in all 16 experimental conditions were recovered with high accuracy. Such high accuracy across all conditions showed that the recovered parameters were unbiased. Second, the subdividing method produced consistent results for the recovered variance components in both rating plans suggesting that as the *rater pool size* increased, the rater-related variance components were estimated with higher precision. However, the second finding held true only if the universe of generalization (Brennan, 1992) matched the patterns and amounts of missing data, which collectively were called a rating plan.

Specifically, if thesparseness of a data set exhibited a pattern resembling that appeared in a *connected mixture* rating plan, only variance components in the $p \times R : I$ universe of generalization can be estimated consistently (i.e. $\hat{\sigma}^2_p, \hat{\sigma}^2_r, \hat{\sigma}^2_{pr}, \hat{\sigma}^2_{ir}, \hat{\sigma}^2_{pir,e}$). Variance components (i.e., $\hat{\sigma}^2_r$, $\hat{\sigma}^2_{pr}$, $\hat{\sigma}^2_{ir}$, $\hat{\sigma}^2_{pir,e}$) in another universe of generalization, namely $p \times I \times R$, were not estimated consistently. When these estimates ($\hat{\sigma}^2_r$, $\hat{\sigma}^2_{pr}$, $\hat{\sigma}^2_{ir}$, $\hat{\sigma}^2_{pir,e}$) were compared to those obtained using the *disconnected crossed* rating plan, they had a lower precision indicating that the *connected mixture* rating plan is frequently less favorable in making decisions for a $p \times I \times R$ universe of generalization. However, results starting on page 26 supported that, in exceptional conditions, the *connected mixture* rating plan can be used to collect data for estimating more precise variance components than those estimated using a *disconnected crossed* rating plan.

The following four sections discussed in details the results corresponding to the four research questions. The performance of the subdividing method was compared and contrasted for the two rating plans in all four sections. In summary, both convergent and discriminant evidences emerged to support

that the subdividing method functioned as expectedly well. That is, the standard error of the variance components that were unrelated to the *person* or the *rater* effects were not influenced by the increment of *sample size* and *rater pool size*, whereas those that were related to the *person* or the *rater* effects were. Tables 4a and 4b delineate an overview of these results for the two rating plans.

Table 4: Effect of increment in *rater pool size* and *sample size* on the standard error of variance components

Table 4A: *Disconnected crossed* rating plan (high standardization)

	$SE(\hat{\sigma}^2_p)$	$SE(\hat{\sigma}^2_i)$	$SE(\hat{\sigma}^2_r)$	$SE(\hat{\sigma}^2_{pi})$	$SE(\hat{\sigma}^2_{pr})$	$SE(\hat{\sigma}^2_{ir})$	$SE(\hat{\sigma}^2_{pir,e})$
Rater pool size			√		√	√	√
Sample size	√			√	√		√

Table 4B: *Connected mixture* rating plan (low standardization)

	$SE(\hat{\sigma}^2_p)$	$SE(\hat{\sigma}^2_i)$	$SE(\hat{\sigma}^2_{pi})$	$SE(\hat{\sigma}^2_{ir})$	$SE(\hat{\sigma}^2_{pir,e})$
Rater pool size				√	√
Sample size	√		√		√

Note. Tables 4A and B. The check marks indicate the variance components influenced by the *rater pool size* and *sample size*.

Q1. How did the amounts and patterns of missing data influence the accuracy and precision of the variance components?

Accuracy

All the variance components were recovered with high accuracy for both the *disconnected crossed* and *connected mixture* rating plans in all 88 experimental conditions, which were composed of the high and low magnitudes of σ^2_i and σ^2_{pr}; large and small *batch size*; high, medium, and low *rater pool sizes*; and the small, medium, large, very large *sample sizes*. For the *disconnected crossed* rating plan, the mean accuracy across the 88 conditions was 100% ± 0.8%. The mean accuracy for making norm-referenced decisions (using the generalizability coefficient) and criterion-referenced decisions (using the dependability coefficient) were 99.8% and 100.5%, respectively (see Table 5). For the *connected mixture* rating plan, the mean accuracy of recovering the five variance component estimates,

shown in Table 6, was 100% ± 0.5%. The mean accuracy of the generalizability and dependability coefficients were 99.9% and 100.6%, respectively.

Table 5: Accuracy of the *disconnected crossed* rating plan

		Min	Mean	Max	Std	n conditions
Accuracy of the Variance Components	person	0.982	0.999	1.026	0.006	88
	item	0.685	0.992	1.251	0.142	88
	rater	0.682	0.993	1.291	0.126	88
	person-by-item	0.986	1.002	1.013	0.005	88
	person-by-rater	0.874	1.002	1.119	0.040	88
	item-by-rater	0.811	1.008	1.431	0.079	88
	person-by-item-rater, error	0.991	1.001	1.011	0.004	88
						88
Accuracy of Composite Indices	Generalizability Coefficient	0.989	0.998	1.005	0.003	88
	Dependability Coefficient	0.957	1.005	1.049	0.011	88

Table 6: Accuracy of the *connected mixture* rating plan

		Min	Mean	Max	Std	n conditions
Accuracy of Variance Components	person	0.984	1.001	1.017	0.006	88
	item	0.679	1.005	1.374	0.169	88
	rater: item	0.825	0.995	1.098	0.053	88
	person-by-item	0.987	1.001	1.014	0.005	88
	person-by-rater-nested-within-item, error	0.993	1.001	1.009	0.003	88
Accuracy of Composite Indices	Generalizability Coefficient	0.989	0.999	1.008	0.003	88
	Dependability Coefficient	0.986	1.006	1.031	0.011	88

As reported in Tables 5 and 6, the accuracy of the *person* and *person-by-item* effects were very similar between the two rating plans, in terms of minimum, mean, maximum, and standard deviations. Although the accuracy of the *item* effects in both rating plans had the largest standard deviations among all other effects, the means of these two accuracy indices tended to converge to one (i.e., 100% accuracy).

BEST COPY AVAILABLE

Such convergence suggested that 100 replications was insufficient to assess the accuracy (or unbiasedness) of the *item* effect. Collapsing across the 88 conditions, however, increased the number of replications to 88,000 and thus provided ample replications to evaluate the highly variable *item* effect, which appeared to be accurately estimated by the subdividing method. Table 5 and Table 6 show that the mean accuracy for *item* effects were, respectively, 99.2% and 100.5% for the *disconnected crossed* and *connected mixture* rating plans.

Standard Errors

High standardization / Disconnected crossed rating plan — This rating plan produced precise estimates for rater-related effects in the universe of generalization exhibiting a $p \times I \times R$ design. Figure 1 (p. 18) summarizes the values of $SE(\hat{\sigma}^2_r)$, corresponding to the *disconnected crossed* rating plan for all the levels of *sample size*, *rater pool*, *batch size*, *item* effect, and *person-by-rater* effect. The precision of the *rater* effect varied only minimally among the different levels of *item* effect, *person-by-rater* effect, and *batch size*. The major variation was due to the size of rater pool; the $SE(\hat{\sigma}^2_r)$ decreased considerably as the *rater pool* expanded, holding the sample size constant. The percent decrease in standard error was lower than one half of the percent increase in the rater pool size. The two horizontal lines, for high and low *person-by-rater* effect conditions (σ^2_{pr} = 0.1 and 0.01), representing the theoretical standard errors for σ^2_r coincided at 0.0131. Even though the theoretical and empirical standard errors were supposed to be the same or at least very close in the conditions where there was no missing data (i.e., 2-rater-750-examinee conditions), the empirical standard errors appeared to be larger than the theoretical standard errors. This finding was not surprising. Rather, it reflected the inaccuracy of the theoretical standard errors based on a small number of levels in a facet (see page 8 for discussions).

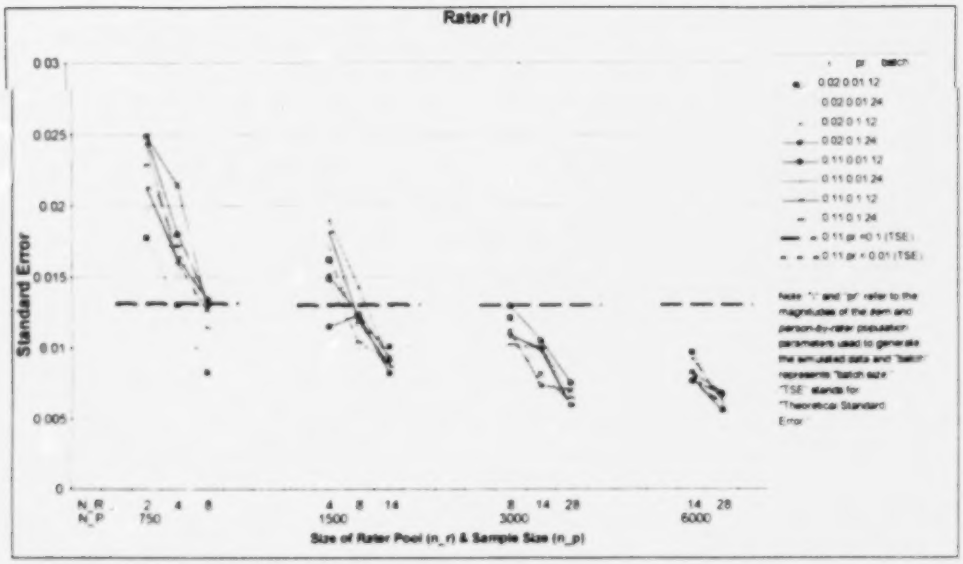


Figure 1: The reduction of standard error for the rater effect as a function of the size of rater pool and sample size (*disconnected crossed* rating plan)

Table 7 displays the average SEs and average reductions of standard error in percentage terms. As the *rater pool size* increased from 2 to 4 (a 100% increase), the standard error declined from 0.0224 to 0.0163 with a 27% reduction, averaged across *item effect*, *person-by-rater effect*, and *batch size*. While holding the *rater pool size* constant at any value, the increase in sample size did not lead to a sizeable decrease in standard error. For instance, given a pool of eight raters, the range of $SE(\hat{\sigma}_r^2)$ was 0.0115 ± 0.0005, despite sample size increase from 750, to 1500, and to 3000. The *item-by-rater* effect showed similar results.

Table 7: Average SEs and average reduction in empirical standard error for the *rater* effect (*disconnected crossed* rating plan)

$\sigma(\delta^2_r)$	Size of Rater Pool 2	4	8	14	28
Sample Size 750	0.0224	27% → 0.0163	27% → 0.0119		
1500		0.0157	25% → 0.0117	25% → 0.0088	
3000			0.0110	18% → 0.0090	31% → 0.0062
6000				0.0083	27% → 0.0060

BEST COPY AVAILABLE

Figure 2 shows standard errors of the *item-by-rater* effect. The trends in the SEs resembled those for the *rater* effect, indicating that expanding the size of a rater pool reduced the standard error of the *item-by-rater* effect.

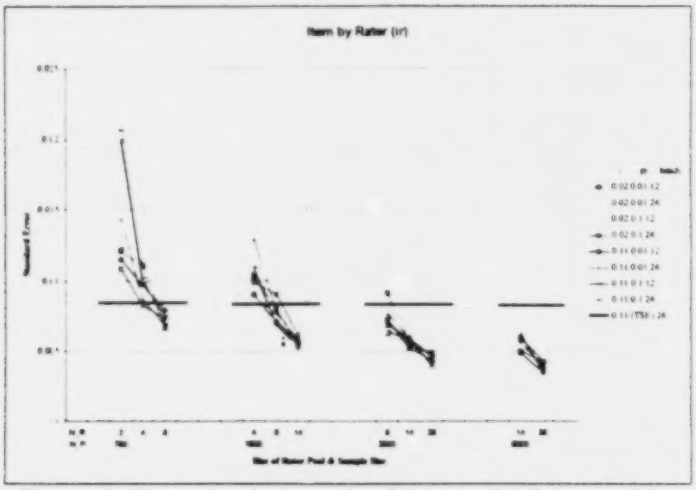


Figure 2: The reduction trends of the standard error of the *rater-by-item* effect

Table 8 shows that the percent reduction in standard error was less than one half of the percent increase in the size of rater pool for the *item-by-rater* effect.

Table 8: Relationship between size of rater pool and reduction in standard error of the *item-by-rater* effect as a function of sample size

$\sigma(\hat{\sigma}^2_{ir})$	Size of Rater Pool				
Sample Size	2	4	8	14	28
750	0.0145	32% 0.0099	25% 0.0074		
1500		0.0100	24% 0.0076	25% 0.0057	
3000			0.0072	20% 0.0058	24% 0.0044
6000				0.0055	29% 0.0039

BEST COPY AVAILABLE

Low standardization / Connected mixture rating plan — The *connected mixture* rating plan *produced precise estimates for the rater-related effect in the universe of generalization exhibiting a* $p \times R : I$ design. The $p \times R : I$ universe of generalization contained only five variance components. These variance components have the following relationship with those in the $p \times R \times I$ universe of generalization (completely crossed universe of generalization). Four of the seven variance components obtained in the completely crossed universe of generalization can be summed to parallel the five variance components, namely $\hat{\sigma}^2_{r,i} = \hat{\sigma}^2_r + \hat{\sigma}^2_{ir}$ and $\hat{\sigma}^2_{pr,i,e} = \hat{\sigma}^2_{pr} + \hat{\sigma}^2_{pir,e}$ (Brennan, 1992). Figure 3 presents the precision of $\hat{\sigma}^2_{r,i}$ obtained from the reconfiguration.

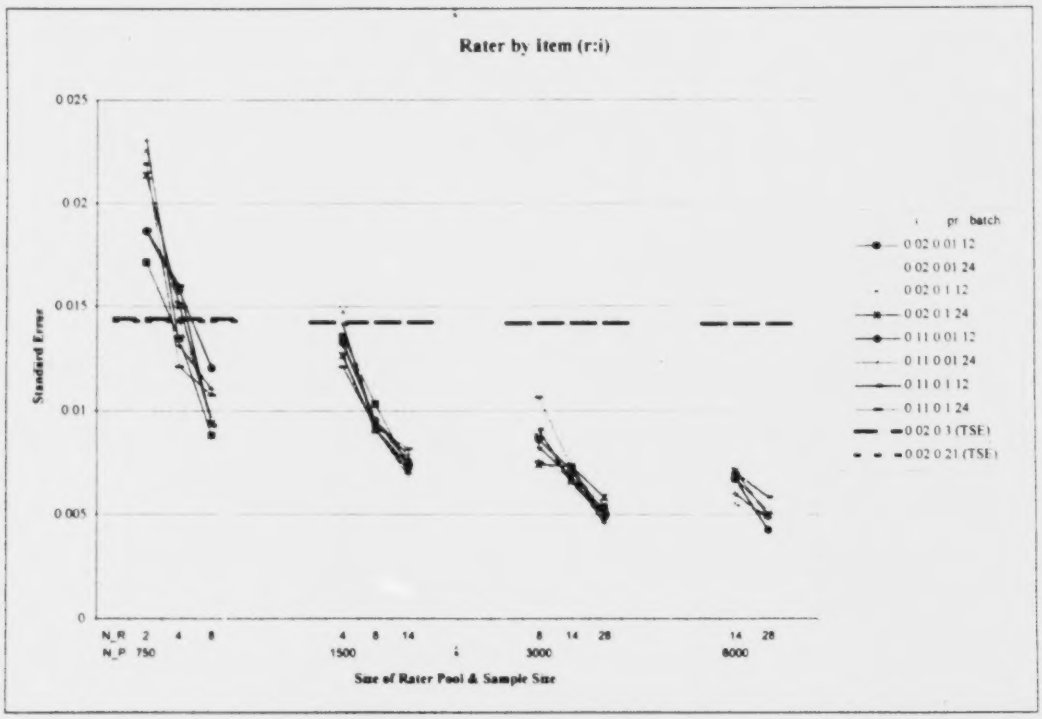


Figure 3: The decrease in standard error as a function of *rater pool size* after utilizing all the available data (connected mixture rating plan)

The average standard error for the 750-examinee conditions (averaged across the *item* effect, *person-by-item* effect, and *batch size*) decreased from 0.0205 to 0.0141, and went further down to 0.0103 yielding a reduction trend of 31% and 27%. The average standard error for the 1500-examinee conditions

BEST COPY AVAILABLE

showed a similar reduction rate with a 30% (0.0135 to 0.0093) decline from sampling 2 of 4 raters to sampling 2 of 8 raters and with a 20% reduction from the 8-rater conditions to the 14-rater conditions. For the 3000-examinee conditions, the decreases were 26% (0.0086 to 0.0068) and 25% (0.0068 to 0.0051), respectively, for the 8-to14-rater expansion and the 14-to-28-rater expansion. Increasing from a pool of 14 raters to 28 raters reduced the average standard error by 23% (0.0065 to 0.005) for a sample of 6000 examinees.

Q2. Can the subdividing method handle a large volume of examinees? How well does it perform?

High standardization / Disconnected cross rating plan — Figures 4 and 5 (p. 22) depict the standard error of the object of measurement $\hat{\sigma}_p^2$ and the *person-by-item-by-rater* (plus systematic and non-systematic errors) effect $\hat{\sigma}_{pir,e}^2$ as a function of *sample size*, *rater pool size*, and three other simulation parameters. For $\hat{\sigma}_p^2$, the two horizontal reference lines represent the theoretical standard errors, given two raters. The top and bottom reference lines reflect standard errors for a large and a small *person-by-rater* effect. The empirical standard errors clustered closely around the two theoretical referenced lines and the standard errors exhibited a pattern similar to that was found for the *person-by-rater* effect. The values of $SE(\hat{\sigma}_e^2)$ reduced by less than 50% as the sample size doubled. Expanding the size of rater pool had inconsistent, thus ignorable effects on $SE(\hat{\sigma}_e^2)$. The standard errors of $\hat{\sigma}_{pir,e}^2$ resembled those of $\hat{\sigma}_p^2$ — increase in the rater pool size had a marginal effect on the reduction of standard error whereas increase in sample size had a significant effect.

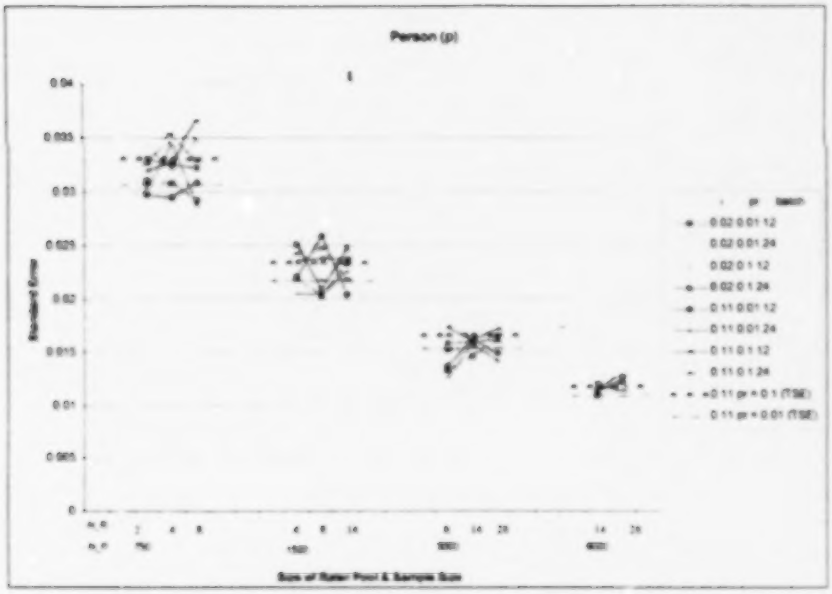


Figure 4: The standard error of the *person* effect as a function of *sample size* and *rater pool size*

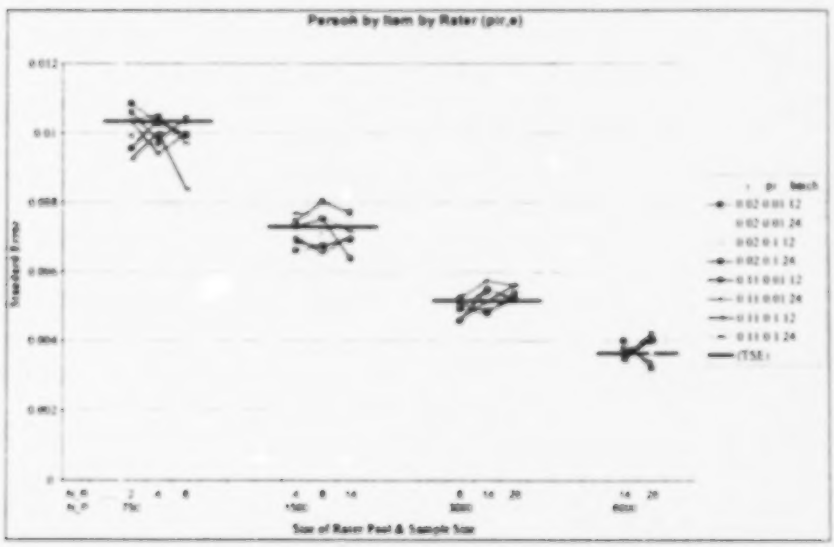


Figure 5: The standard error of the *person-by-item-by-rater* effect as a function of *sample size* and *rater pool size*

BEST COPY AVAILABLE

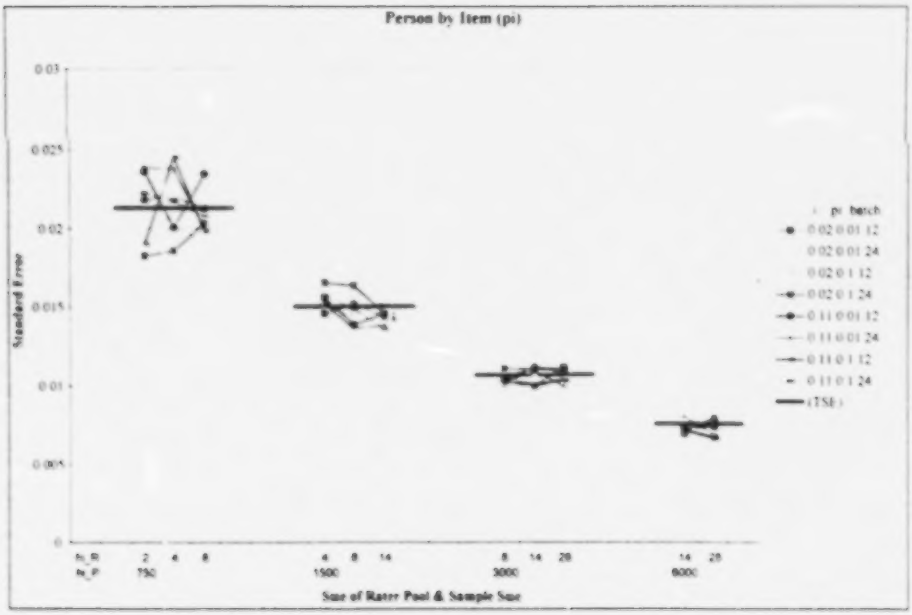


Figure 6: The standard error of the *person-by-item* effect as a function of sample size

Figure 6 depicted the standard errors of the *person-by-item* effect and it shows results coherent with those for the previous three variance components $\hat{\sigma}^2_p$, $\hat{\sigma}^2_{pr}$, and $\hat{\sigma}^2_{pr,e}$. Specifically, the statistical property of consistency (the larger the sample size the smaller the variability) holds true for the subdividing method.

Figure 7 on page 24 shows values of the standard error of the *person-by-rater* effect. The standard error decreased as the sample size increased, holding constant the size of the rater pool. Expanding the rater pool did not increase precision for the *person-by-rater* effect. The two series of trends clustered around the two reference lines for the two levels of theoretical standard errors, indicating that the $SE(\hat{\sigma}^2_{pr})$ obtained by subsampling was comparable to that obtained by employing two raters regardless of the size of the rater pool. Table 9 on page 24 shows the percent reduction in the SEs as a function of the sample size given a small *person-by-rater* effect (σ^2_{pr} =0.01). The percentages of standard error reduction for all the conditions were less than the percent of increase in sample size and all

BEST COPY AVAILABLE

increments to sample size reduced the standard errors. This observation also applied for a large *person-by-rater* effect (σ^2_{pr} =0.10).

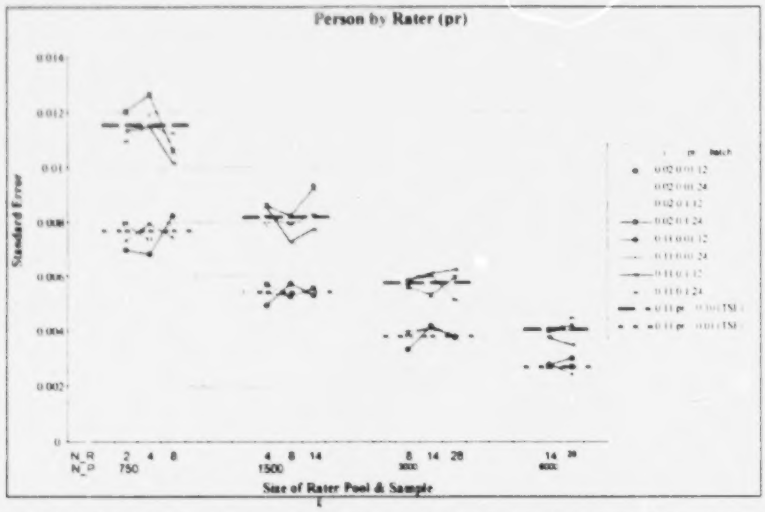


Figure 7: The standard error of the *person-by-rater* effect as a function of sample size

Table 9: SE and changes in standard error of the *person-by-rater* effect as *sample size* increases

$\sigma(\hat{\sigma}^2_{pr})$	Size of Rater Pool				
	2	4	8	14	28
Sample Size 750	0.0075	27% 0.0074	30% 0.0079		
1500		0.0053	32% 0.0055	25% 0.0055	
3000			0.0037	0.0041	31% 0.0039
6000				34% 0.0027	0.0027

Low standardization / Connected mixture rating plan — Results of the variance components obtained for the *connected mixture* rating plan resembled the results of the *disconnected crossed* rating plan. Figures 8a-c show values of the standard errors of the *person*, *person-by-item*, and the *person-by-rater-nested-in-item* effect. The standard errors decreased as sample size increased, holding constant the size of the rater pool. Expanding the rater pool did not increase precision for the three effects.

BEST COPY AVAILABLE

Figure 8a: Standard error

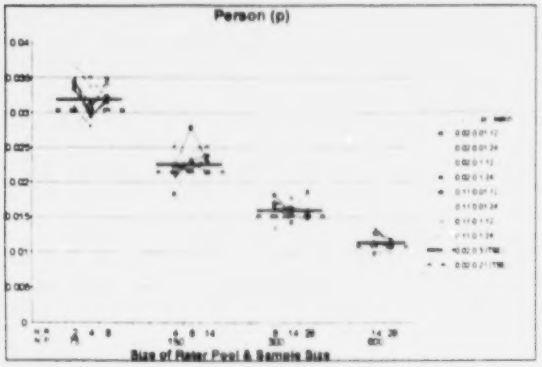


Figure 8b: Standard error

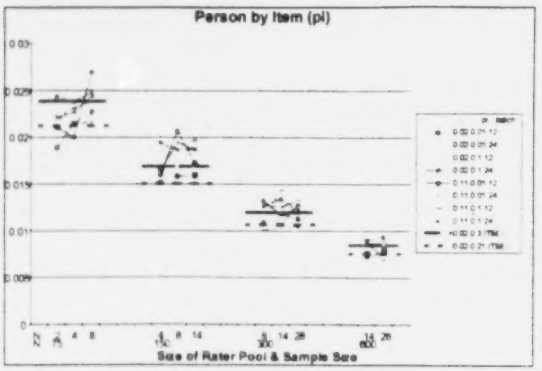
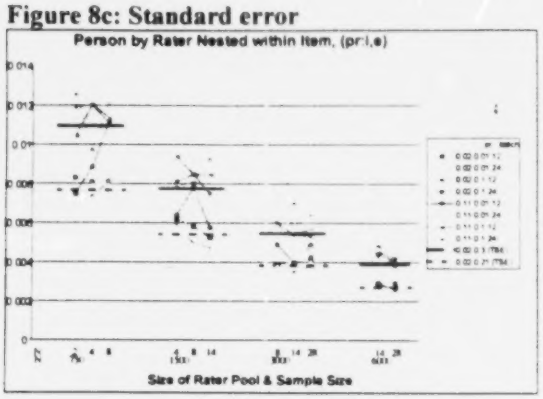


Figure 8c: Standard error



Q3. To what degree does increasing the number of essays scored by a common group of raters increase the precision of the variance component estimates and the reliability coefficient estimates?

Packing essays into batches of 24 as opposed to 12 did not systematically raise nor lower the accuracy and precision of variance components and composite indices in both rating plans. The mean ratios of the standard errors across the 44 conditions for each of the variance components were very close to one (± 0.05). This result indicated that the average estimates obtained by packing essays into batches of 24 were as precise as those obtained by packing essays into batches of 12. Consistent results were found for both rating plans with high and low standardizations. Tables 10 and 11 (p. 26) show the ratios for the variance components in the two rating plans.

BEST COPY AVAILABLE

Table 10: High standardization: disconnected crossed rating plan

		Min	Mean	Max	Std	n conditions
SE Ratios for Variance Components	person	0.790	**0.996**	1.185	0.085	44
	item	0.558	**1.036**	1.639	0.272	44
	rater	0.748	**1.030**	1.660	0.176	44
	person-by-item	0.824	**1.007**	1.187	0.086	44
	person-by-rater	0.806	**1.018**	1.281	0.095	44
	item-by-rater	0.724	**1.016**	1.430	0.156	44
	person-by-item-rater, error	0.836	**1.037**	1.173	0.086	44

Table 11: Low standardization: *connected mixture* rating plan

		Min	Mean	Max	Std	n conditions
SE Ratios for Variance Components	person	0.733	**0.997**	1.157	0.090	44
	item	0.570	**1.009**	1.767	0.293	44
	rater	0.757	**1.019**	1.358	0.122	44
	person-by-item	0.763	**0.995**	1.281	0.124	44
	person-by-rater-nested-in-item, error	0.638	**0.947**	1.299	0.138	44

Q4. When should rating plans with high standardization be used in placed of rating plans with low standardization?

Table 12 (p. 27) summarizes that the expansion of the rater pool reduced the certainty of the *person-by-rater* effect. It can be observed that all the increase in imprecision (or increase in SEs) in Table 12 (indicated by arrows) were less than approximately five percent in addition to half of the percentage increase in the *rater pool size*. Expanding the *rater pool size* from four to eight raters yielded a 100% increase in rater pool size and the increases in uncertainty for the *person-by-rater* effect were 49% and 54%, respectively, for sample sizes 750 and 1500. The corresponding increases in uncertainty became

BEST COPY AVAILABLE

lower (21% and 31% for sample sizes 1,500 and 3,000) as the rater pool expanded by a lower percentage, 75% as opposed to 100% (from eight to 14 raters). The increases in uncertainty returned to the mid- and high- forty percent (47% and 44%) when the rater pool expanded by 100%, for sample sizes 3,000 and 6,000. All of the above reductions were less than 55% (i.e., 0.5 * 100% + 5%).

Table 12: Increases in uncertainty of the *person-by-rater* effect in the *connected mixture* rating plan

$\sigma(\hat{\sigma}^2_{pr})$	4	8 (100%)	14 (75%)	28 (100%)
Sample Size 750	0.0226	0.0338 (49%)		
Sample Size 1500	0.0162	0.0249 (54%)	0.0302 (21%)	
Sample Size 3000		0.0168	0.0219 (31%)	0.0323 (47%)
Sample Size 6000			0.0160	0.0230 (44%)

Figure 9 (p.28) depicts the phenomenon reported in Table 12 accompanied by the theoretical standard errors. These theoretical standard errors predicted the *person-by-rater* effect based on the same two raters scoring all the examinees on all items (i.e., completely balanced situations with a crossed structural design).

BEST COPY AVAILABLE

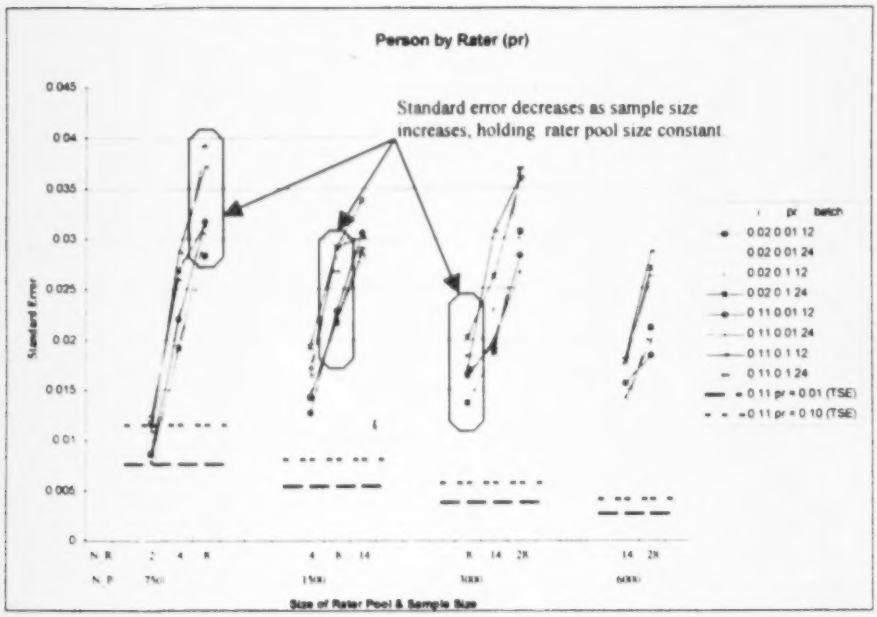


Figure 9: The relationship between the improvement of the *person-by-rater* effect and the expansion of sample size using the *connected mixture* rating plan.

As observed in the above figure, the larger the rater pool, the farther was the empirical standard error from the theoretical standard error. This indicates that one would become less confident of the *person-by-rater* effect as more raters were employed, holding the number of examinees constant.

The increase in sample size had an opposite effect than that of expanding the size of rater pool — holding the size of rater pool constant, the larger the sample size, the higher the precision (and the smaller the SEs). Such observations applied to all four levels of sample size. The trend shows the increase in precision as the sample size increased from 750, 1500, to 3000, holding constant the size of the rater pool (the three boxes in Figure 9 highlight the 8-rater pool examples; the larger the sample size, the smaller the SEs). The means of $SE(\hat{\sigma}^2_{pr})$ were 0.0315, 0.0249, and 0.0144 for the three levels of sample sizes (750, 1500, and 3000). The minimum reduction in $SE(\hat{\sigma}^2_{pr})$, from one level of sample size to another, was over 25%.

Figure 10 on page 29 compares the degree of uncertainty of the *person-by-rater* effect in the *connected mixture* rating plan to that in the *disconnected crossed* rating plan ($SE(\hat{\sigma}^2_{pr}) - SE(\hat{\sigma}^2_{pr})$), where

BEST COPY AVAILABLE

$\hat{\sigma}^2_{pr}$ and $\hat{\sigma}^2_{pr'}$ represented the *person-by-rater* effects in the *disconnected crossed* and *connected mixture* rating plans, respectively. The $\hat{\sigma}^2_{pr}$ was estimated with a higher degree of precision in the *disconnected crossed* rating plan. On average across all the conditions for the 750-examinee conditions, $\hat{\sigma}^2_{pr}$ was estimated with a precision 0.0127 higher than it was estimated in the *connected crossed* rating plan. For the 1500-, 3000-, and 6000- examinee conditions, $\hat{\sigma}^2_{pr}$ was estimated with even higher precision: mean differences were 0.0170, 0.0189, and 0.0161, respectively. With no missing data appeared (*rater pool size* = 2), the difference in standard errors between the two rating plan was almost zero (mean difference = -0.0005) implying that both rating plans had equal precision in estimating the *person-by-rater* effect.

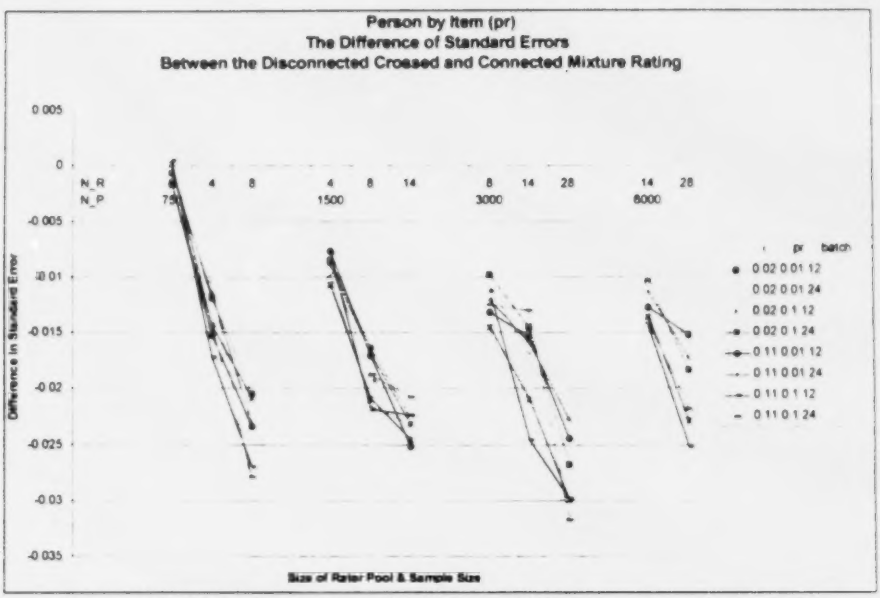


Figure 10: The effect of employing two different rating plans on the precision of the person-by-rater effect

Holding *rater pool size* constant, the difference between the standard errors diminished as the sample size increased. For instance, given a *rater pool* size 8, the average standard errors increased from -0.024 to -0.018, which then increased to -0.012 for the 750-, 1500-, and 3000-examinee conditions. While keeping the same *rater pool size* (8), one can extrapolate a positive difference in standard error if a

BEST COPY AVAILABLE

very large sample size is used. Such a positive difference implies that the c*onnected mixture* rating plan will outperform the *disconnected crossed* rating plan in producing precise variance components for cases where tremendous examinees take a test but they are scored by only relatively small pool of raters (e.g., 6000 examinees and 4 raters).

As was true for $\hat{\sigma}^2_{pr}$, three other rater-related variance components ($\hat{\sigma}^2_{pir,e}$, $\hat{\sigma}^2_{r}$, $\hat{\sigma}^2_{ir}$) manifested the same relationship among *sample size*, *rater pool size*, and the difference in standard errors. Figure 11a to Figure 11d on page 31 display these relationships in a panel. Figure 11b and 11d show that the *disconnected crossed* rating plan did not outperform the *connected mixture* plan as much as it did for the effects involving the *person* effect (i.e., $\hat{\sigma}^2_{pr}$ and $\hat{\sigma}^2_{pir,e}$). For example, given the 3000-examinee-8-rater condition, the *disconnected* crossed plan outperformed the *connected mixture* plan by an average 0.018 unit in estimating $SE(\hat{\sigma}^2_{pr})$ (Figure 11a) but it outperformed the *connected mixture* plan by only 0.002 unit in estimating $SE(\hat{\sigma}^2_{r})$ (Figure 11b).

Figure 11: Difference of Standard Errors

Between the *Disconnected Crossed* and *Connected Mixture* Rating Plans

Figure 11a

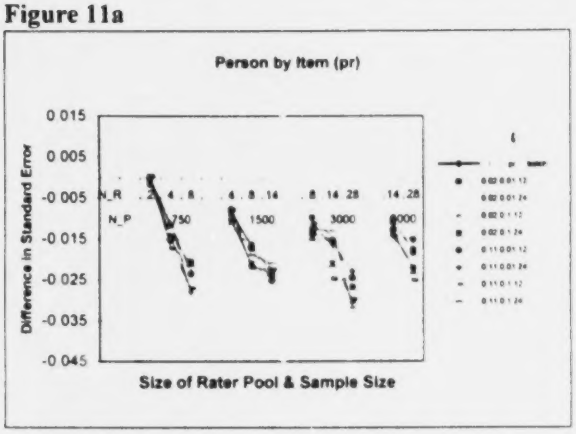


Figure 11b

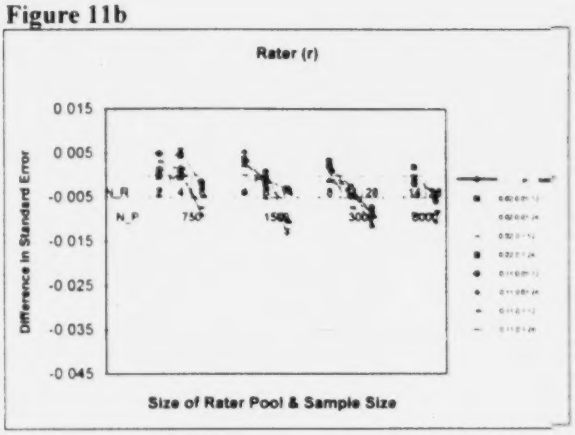


Figure 11c

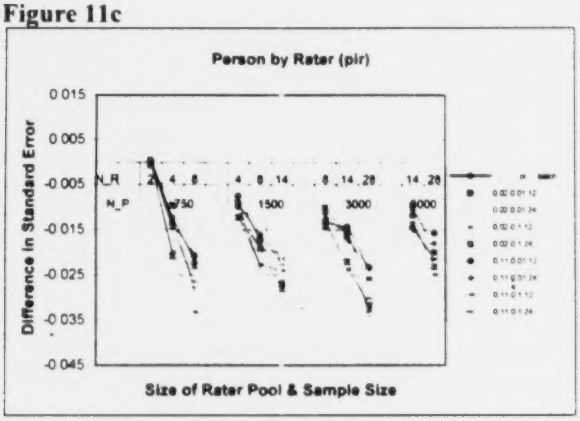
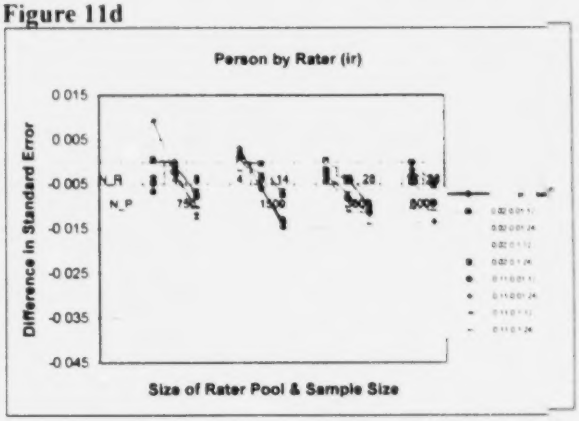


Figure 11d



BEST COPY AVAILABLE

Discussions

Scoring constructed response items is more time consuming and complex than scoring multiple-choice items. Many educational and non-educational institutions adapt open-ended questions in examinations for admission, certification, graduation, accountability, and licensing purposes. These examinations often are administered on a large-scale basis. Large volumes of examinees are tested and yet only a limited time (usually a few weeks) is available for scoring. Hundreds of raters are recruited to score the examinations and it is infeasible to assign all the raters to score every one of the examinees. Thus each examinee will be scored by a selection of raters, leading to sparsely-filled data sets and also unbalanced designs. Sparsely-filled data sets could potentially cause biased and imprecise estimators. The current study developed and validated a method, called the subdividing method, to resolve this problem. The subdividing method, drawing on the concept that one could obtain more stable estimators by synthesizing multiple data sources than using just one source, is set out to improve the accuracy and precision of estimates quantifying measurement errors in the framework of G theory. The estimates produced by the subdividing method were scrutinized in determining how well the method worked in realistic scenarios and these scenarios included differences in: (1) volume of examinees, (2) size of rater pool, (3) variation of item difficulty, (4) levels of rater inconsistency, (5) rules used to decide how to group raters and assign tasks to raters, and (6) the minimum number of examinees scored by a group of raters. Results reported in the study indicated that the subdividing method produced outcomes having properties (unbaisedness and consistency) that are similar to those of complete data methods. For instance, the larger the volume of the examinees, the more precisely the following effects were estimated: *person*, *person-by-item*, *person-by-rater*, and *person-by-item-by-rater* effects. In addition, even though every essay was scored by only two randomly selected raters from a large pool of raters, increasing the size of the pool of raters can improve the precision in determining rater-related effects (i.e., *rater*, *item-by-rater*, *rater-nested-in-item* effects).

Unlike precision, the accuracy of estimating measurement errors did not depend on the amounts and patterns of missing data (i.e., rating plans). Accuracy of the outcomes was very high, close to perfect. These finding substantiated that the subdividing method produced unbiased outcomes with data missing by design.

The subdividing method recovers variance component estimates with high accuracy and precision in a variety of conditions (i.e., low and high variations in item difficulty and rater inconsistency). Increasing the number of examinees scored by the same group of raters from 12 to 24 had virtually no effect on the accuracy and precision of the variance component estimates.

Highly standardized rating plan (*disconnected crossed* rating plan) estimated variance components with accuracy and precision for a universe of generalization exhibiting a $p \times I \times R$ design; whereas rating plans with low standardization (*connected mixture* rating plan) estimated precise and accurate variance components for a universe of generalization exhibiting a $p \times R : I$ design. If test scores were collected using a *connected* mixture rating plan and were used to generalize to a universe of generalization $p \times I \times R$, the rater-related variance components were not estimated with a high precision as they were using high standardization. However, in unusual occasions where only a relatively few raters were used to score a large number of examinees, the *connected mixture* rating plan could outperform the *disconnected mixture* plan in producing precise variance component estimates.

Different rules used for forming groups of raters changed the structural design of scores and thus influenced the precision of measurement error estimation. The results from a Monte Carlo study concluded that scoring centers can determine and forecast the confidence they can put in the generalizability analyses by controlling (1) the rules used to assign tasks to raters and (2) the universe of generalization for which normative and criterion-referenced decisions are made. Suggestions are provided in the study for the set-up of scoring procedures.

Educational values

Given the proliferation of performance assessments, many states and school districts have already implemented this type of assessment on a regular basis. Well-developed scoring rubrics can be useful, provided that raters implement them consistently and accurately. Large-scale performance-based assessments can be used for accountability purposes only if methods are developed to evaluate the quality of the scoring procedures (Mehrens,1992). Training raters to consistently apply the scoring criteria described in the rubrics designed for large-scale performance assessments, has many instructional benefits. When raters (mostly school teachers) return to their classrooms, they will be accustomed to using those criteria.

Improving the quality of performance assessment so that it can be used for high-stakes decisions also can help align assessments with curriculum and instruction (Pearson, 1998). Many researchers such as Bracey (1989) reported that schools and teachers were less likely to include materials in their classrooms if the materials would not be tested in high-stakes examinations. Developing methods to monitor and improve the quality of performance assessment could reduce the tensions in using this state-of-the-art assessment for classroom instruction and for high-stakes decisions. Students will be the beneficiaries of this development, which was examined in the current study.

Appendix A:

The structure of a *modified balanced incomplete block* design

The *MBIB*, a variation of the balanced incomplete block design (BIB) (e.g., Montgomery, 1997, p.208; Searle et al., 1992, p.5), is a balanced design because every examinee receives two ratings on both essays. The incompleteness comes into play due to the fact that one reader grades the essay twice and yet the two other different raters grade the essay only once. The *MBIB* design is specially designed for situations differing from the BIB design that the *MBIB* design has an additional factor to the BIB design. As can be seen in Figure 12 there are two levels in the item factors (items 1 and 2). As can be observed in Figure 12 a data subset with the *MBIB* design using raters A, B, and C, with A being the rater assign scores to both items, can have n_{p31} examinees. The first subscript p indicates that the sample size n refers to the number of examinees. The second subscript 3 indicates that the data set exhibits a *MBIB* design and the third subscript 1 indicates that the data set is the first set in the *MBIB* design. The total number of cases that can be analyzed by the *MBIB* model is $n_{p3\cdot} = \sum_{x=1}^{3} n_{p1x}$. The entire sparsely filled data has a total number of cases of $N = \sum_{index=1}^{3} \sum_{x=1}^{3} n_{p1x}$.

Figure 12: Hypothetical data subsets for the *modified balanced incomplete block* (MBIB) design

	Item 1			Item 2		
	Rater A	Rater B	Rater C	Rater A	Rater B	Rater C
Examinee 1	$X_{3,1,1,1,A}$	$X_{3,1,1,1,B}$		$X_{3,1,1,2,A}$		$X_{3,1,1,2,C}$
Examinee 2	$X_{3,2,1,1,A}$		$X_{3,2,1,1,C}$		$X_{3,2,1,2,B}$	$X_{3,2,1,2,C}$
Examinee 3		$X_{3,3,1,1,B}$	$X_{3,3,1,1,C}$	$X_{3,3,1,2,A}$	$X_{3,3,1,2,B}$	
Examinee 4	$X_{3,4,1,1,A}$	$X_{3,4,1,1,B}$			$X_{3,4,1,2,B}$	$X_{3,4,1,2,C}$
Examinee 5	$X_{3,5,1,1,A}$		$X_{3,5,1,1,C}$	$X_{3,5,1,2,A}$	$X_{3,5,1,2,B}$	
Examinee 6		$X_{3,6,1,1,B}$	$X_{3,6,1,1,C}$	$X_{3,6,1,2,A}$		$X_{3,6,1,2,C}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Examinee m	$X_{3,St,Sp,2*1,A}$	$X_{3,St,Sp,2*1,B}$			$X_{3,St,Sp,2*2,B}$	$X_{3,St,Sp,2*2,C}$

Note. The Xs indicate scores assigned to an essay. The subscripts indicate the location of the score, where location is defined as the structural design (first subscript), data subset (second subscript), examinee (third subscript), item (fourth subscript), and rater (fifth subscript) for the corresponding score.

BEST COPY AVAILABLE

Appendix B:

A mathematical model to determine the size of a rater pool

To figure the number of raters needed for the entire scoring procedure, we first develop a simple mathematical model. Equation (4) gives a generic model where the number of raters needed is determined by two quantities, namely the total number of ratings in an administration of a test (denoted *Total workload in minutes*) and the total amount of time, in minutes, that an average rater can put in during the entire scoring procedure (denoted *Total work time per rater*).

$$Number_of_raters_needed = Total_workload_in_minutes / Total_work_time_per_rater \quad (4)$$

As can be observed, more raters are needed as the total workload increases while holding the total work time per rater constant. By the same token, fewer raters are needed as the total workload decreases.

We now define more specifically *Total workload in minutes* and *Total work time per rater* in terms of other practical constraints such as the available time for generating and reporting the scores. By defining those two quantities as a function of other practical constraints, we will be able to decide how many raters are needed given the available resources (e.g., how much time do we have until the scores must be analyzed and reported?). The definitions of those two quantities can be modified to accommodate constraints of individual scoring centers. In addition, the values of those two quantities can vary from one administration to the others within an individual scoring center, depending upon policy, needs, and available resources. In the examples that follow, we define the total workload and total work time in terms of constraints common to essay scoring, according to anecdotal reports. Thus,

$$Total_workload_in_minutes = n_examinees \times n_essays_answered_by_an_examinee \times n_ratings_per_essay \times reading_time_per_essay_in_minutes, \text{ and} \quad (5)$$

$$Total_work_time_per_rater = n_total_scoring_days \times n_work_hours_per_rater_per_day \times 60_minutes. \quad (6)$$

By substituting values into Equations (5) and (6) into (4), we reach the following symbolic equation.

$$\text{Number of raters needed} = \frac{n_p \bullet n_i \bullet n_{r/e} \bullet t_e}{n_d \bullet n_h \bullet 60}, \qquad (7)$$

where n_p = number of examinees,

n_i = number of items (essays) responded to by an examinee,

$n_{r/e}$ = number of ratings on each essay,

t_e = average time needed to score an essay by a rater (in minute),

n_d = number of days available to complete scoring, and

n_h = average work hours per day by a rater.

For instance, let us assume 6,000 examinees took a test and each examinee responded to two 500-word writing prompts. Two different raters score each writing prompt. We further assume that it takes an average of 10 minutes for a reading of an essay. Substituting this information into Equation (7), we need 14 raters to complete the scoring in 40 workdays of 7.5 hours ((6000*2*2*10)/(40*7.5*60) = 13.3).

Appendix C:

Standard error for variance components in a two facet crossed design[1]

$$\sigma(\hat{\sigma}^2_p) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_i \cdot \hat{\sigma}^2_{pi} + n_r \cdot \hat{\sigma}^2_{pr} + \cdot n_i \cdot n_r \cdot \hat{\sigma}^2_p)^2}{(n_p+1) \cdot n_i^2 \cdot n_r^2} + 2\frac{(\hat{\sigma}^2_{pir,e} + n_i \cdot \hat{\sigma}^2_{pi})^2}{((n_p-1) \cdot (n_i-1)+2) \cdot n_i^2 \cdot n_r^2} \right. \quad (8)$$
$$\left. + 2\frac{(\hat{\sigma}^2_{pir,e} + n_i \cdot \hat{\sigma}^2_{pr})^2}{((n_p-1) \cdot (n_r-1)+2) \cdot n_i^2 \cdot n_r^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1) \cdot (n_r-1)+2) \cdot n_i^2 \cdot n_r^2} \right]^{\frac{1}{2}}$$

$$\sigma(\hat{\sigma}^2_i) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_p \cdot \hat{\sigma}^2_{ir} + n_r \cdot \hat{\sigma}^2_{pi} + \cdot n_p \cdot n_r \cdot \hat{\sigma}^2_i)^2}{(n_i+1) \cdot n_p^2 \cdot n_r^2} + 2\frac{(\hat{\sigma}^2_{pir,e} + n_r \cdot \hat{\sigma}^2_{pi})^2}{((n_p-1) \cdot (n_i-1)+2) \cdot n_p^2 \cdot n_r^2} \right. \quad (9)$$
$$\left. + 2\frac{(\hat{\sigma}^2_{pir,e} + n_p \cdot \hat{\sigma}^2_{ir})^2}{((n_i-1) \cdot (n_r-1)+2) \cdot n_p^2 \cdot n_r^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1) \cdot (n_r-1)+2) \cdot n_p^2 \cdot n_r^2} \right]^{\frac{1}{2}}$$

$$\sigma(\hat{\sigma}^2_r) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_p \cdot \hat{\sigma}^2_{ir} + n_i \cdot \hat{\sigma}^2_{pr} + \cdot n_p n_i \cdot \hat{\sigma}^2_r)^2}{(n_r+1) \cdot n_i^2 \cdot n_p^2} + 2\frac{(\hat{\sigma}^2_{pir,e} + n_i \cdot \hat{\sigma}^2_{pr})^2}{((n_p-1) \cdot (n_r-1)+2) \cdot n_i^2 \cdot n_p^2} \right. \quad (10)$$
$$\left. + 2\frac{(\hat{\sigma}^2_{pir,e} + n_p \cdot \hat{\sigma}^2_{ir})^2}{((n_i-1) \cdot (n_r-1)+2) \cdot n_i^2 \cdot n_p^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1) \cdot (n_r-1)+2) \cdot n_i^2 \cdot n_p^2} \right]^{\frac{1}{2}}$$

$$\sigma(\hat{\sigma}^2_{pi}) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_r \cdot \hat{\sigma}^2_{pi})^2}{((n_p-1) \cdot (n_i-1)+2) \cdot n_r^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1)(n_r-1)+2) \cdot n_r^2} \right]^{\frac{1}{2}} \quad (11)$$

$$\sigma(\hat{\sigma}^2_{pr}) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_i \cdot \hat{\sigma}^2_{pr})^2}{((n_p-1) \cdot (n_r-1)+2) \cdot n_i^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1)(n_r-1)+2) \cdot n_i^2} \right]^{\frac{1}{2}} \quad (12)$$

$$\sigma(\hat{\sigma}^2_{ir}) = \left[2\frac{(\hat{\sigma}^2_{pir,e} + n_p \cdot \hat{\sigma}^2_{ir})^2}{((n_p-1) \cdot (n_r-1)+2) \cdot n_p^2} + 2\frac{(\hat{\sigma}^2_{pir,e})^2}{((n_p-1) \cdot (n_i-1)(n_r-1)+2) \cdot n_p^2} \right]^{\frac{1}{2}} \quad (13)$$

$$\sigma(\hat{\sigma}^2_{pir,e}) = \left[2\frac{(\hat{\sigma}^2_{pir,e})^2}{(n_p-1) \cdot (n_i-1)(n_r-1)+2} \right]^{\frac{1}{2}} \quad (14)$$

[1] Equations for the standard errors were derived by substituting variance components, sample sizes, and mean squares into the general formula reported in Brennen (1992, p. 101, 6.2.1).

BEST COPY AVAILABLE

References

Bell, J. F. (1985). Generalizability theory: The software problem. *Journal of Educational Statistics, 10*(1), 19-29.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City: American College Testing.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14-20.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement, 55*(2), 157-176.

Burdick, R. K. & F. A. Graybill (1992). *Confidence Intervals on Variance Components*. New York, Marcel Dekker.

Chiu, C. W. T. (1999). *Scoring Performance Assessment Based on Judgements: Utilizing Meta-Analysis to Estimate Variance Components in Generalizability Theory for Unbalanced Situations.*. Unpublished Doctoral Dissertation. Michigan State University. E. Lansing, MI.

Chiu, C. W. T., & Wolfe, E. W. (1997, April). *Generalizability Theory: A New Approach to Analyze Non-Crossed Performance Assessment Data.* Paper presented at the American Educational Research Association annual meeting, Chicago, IL.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*(1), 19-93.

Goodnight, J. H. (1978). *Computing MIVQUE0 Estimates of Variance Components* (SAS Techincal Report R-105). Cary, NC: SAS Institute.

Giesbrecht, F. G. (1983). An efficient procedure for computing MINQUE of variance components and generalized least squares estimates of fixed effects. *Communications in Statistics, Series A: Theory and Method, 12*, 2169-2177.

Gordon, B. (1998, Sept). *Scoring performance assessment.* Personal Communication. University of Georgia.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychology Measurement, 20*(2), 101-125.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for Meta-Analysis*. New York: Academic Press.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a Mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Linn, R. L., Burton, E., DeStefano, L., & Hanson, M. (1996). Generalizability of New Standards Project 1993 Pilot Study Tasks in Mathematics. *Applied Measurement in Education, 9*(3), 201-214.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Longford, N. T. (1995). *Models for uncertainty in educational testing*. New York: Spring-Verlag.

Mooney, C. Z. (1997). *Monte Carlo Simulation*. (Vol. no. 116). Thousand Oaks, CA: Sage.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1995). *Reader calibration and its potential role in equating for the Test of Written English* (Report prepared by the Center for Performance Assessment MS # 95-02). Princeton: Educational Testing Service.

Othman, A. R. (1995). *Examining task sampling variability in science performance assessments*. Unpublished Dissertation. University of California, Santa Barbara.

Patz, R. J., Wilson, M., & Hoskens, M. (1997). *Optimal Rating Procedures and Methodology for NAEP Open-ended Items* (Technical Report Number 97-37). Tempa, Florida: National Center for Educational Statistics (NCES).

Psychometrika Editorial Board. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika*, 44(2), 133-4.

Rao, P. S. R. S. (1997). *Variance components estimation: Mixed models, methodologies and applications*. New York, NY: Chapman & Hall.

Schafer, W. (1998, Sept). *Scoring performance assessments*. Personal Communication. The Maryland School Performance Assessment Program.

Schroeder, M. L. (1986). *Inferential Procedures for Multifaceted Coefficients of Generalizability*. Unpublished Dissertation. The University of British Columbia (CANADA).

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. New York, NY: Wiley.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Vickers, D. (1998, Sept). *Scoring performance assessments*. Personal Communication. North Carolina Performance Assessment Program.

Welch, C. (1996, July). *Scoring performance assessments*. Personal Communication. Performance Assessment Center, ACT, Inc.

Wolfe, E. W. (1998, Sept). *Scoring performance assessment*. Personal Communication. University of Florida.